AI in Science: When Measurement Instruments Learn

INDRĖ ŽLIOBAITĖ
Department of Computer Science, University of Helsinki, Helsinki, Finland
E-mail: indre.zliobaite@helsinki.fi

2025 01 18

Abstract

The growing use of artificial intelligence (AI) in scientific research is reshaping how measurements are produced, interpreted, and trusted. Rather than relying on fixed, physically motivated measurement functions and summary statistics, contemporary science increasingly employs computational inference methods that learn mappings from data to scientifically relevant quantities. These learned measurement instruments support the analysis of complex, large-scale data and open new possibilities for scientific discovery, but they also challenge classical assumptions about measurement, uncertainty, and epistemic responsibility. In this paper, I examine AI-based methods in their role as learned measurement instruments. The behaviour of these instruments is determined not only by design and calibration, but also by training data, modeling choices, and implicit assumptions.

I argue that such methods introduce forms of epistemic risk that go beyond stochastic noise and are not fully captured by existing measurement theory or current explainable AI techniques. In particular, I show that when measurement instruments learn, epistemic uncertainty no longer merely constrains inference drawn from measurements, but plays a formative role in determining what counts as a measurement outcome in the first place. At the same time, because these instruments perform inferential tasks that were traditionally part of human scientific reasoning, their use blurs the traditional boundary between instruments and agents, particularly when AI systems are described as generating hypotheses, guiding experimental design, or contributing to scientific reasoning.

By analyzing this reconfiguration of measurement and inference, I clarify what changes when measurement instruments learn and show how AI reshapes the traditional boundary between instruments and agents in science without attributing strong agency to AI systems. I conclude by outlining the epistemic risks and responsibilities that arise from delegating measurement and inference to learned instrumentation, and by reflecting on how scientific understanding can be maintained when using AI-based and partially non-transparent instrumentation for scientific inference.

1. Introduction: measurement, inference and learning instruments

Scientific measurement has not been merely about recording values; it has long involved inference about what those values mean, how they relate, and whether they warrant belief. Measurements become scientifically informative only through inferential work that connects observations to claims about the world. Traditionally, this work has been carried out by human scientists, while instruments have served to implement measurement procedures defined and interpreted within established conceptual frameworks.

Over recent decades, computational methods have become central to scientific measurement and subsequent data analysis practices. Statistical models and automated pipelines have supported inferential reasoning by summarizing data and estimating uncertainty. More recently, artificial intelligence (AI), particularly machine-learning-based systems, have been increasingly integrated into scientific workflows (Wang *et al.*, 2023) in a qualitatively different way. Rather than merely assisting human reasoning, AI-based systems relocate parts of this inferential work into the instruments themselves by learning mappings from data to scientifically meaningful quantities.

Historically, the inferential tasks that transform measurements into scientific evidence were performed by human scientists. These tasks included determining which aspects of data were scientifically relevant, distinguishing signal from noise, relating measured quantities to theoretical constructs, and judging the strength and scope of empirical support for competing claims. While instruments supplied values and observations, scientists decided how those values should be interpreted, which patterns warranted explanation, and how measurement uncertainty should constrain inference.

In the use of learned measurement instruments, some of these inferential tasks are partially delegated to computational systems. AI-based methods extract representations from data, prioritize certain variables or patterns over others, and implicitly structure the space of plausible relations among observables. In doing so, they shape what counts as a meaningful measurement outcome and how evidence is organized, often without explicit theoretical articulation. This shift does not eliminate human judgment, but it changes where and how inferential decisions are made within scientific practice.

When measurement instruments learn, they no longer function merely as implementations of fixed, physically motivated measurement functions. Instead, they participate in the inferential processes that connect data to scientific claims. As a result, measurement outcomes depend not only on experimental design and calibration, but also on training data, modeling choices, and implicit assumptions embedded in computational systems. These dependencies are often opaque to scientific users, particularly when AI-based systems used as measurement instruments are developed by toolmakers who are not domain experts in the phenomena under investigation.

This delegation of inferential work challenges classical assumptions in measurement theory (Tal, 2013) and alters the epistemic role of scientific instruments. It also blurs the traditional boundary between instruments and agents--not because AI systems possess intentions or understanding, but because they perform inferential functions that have traditionally been part of scientific reasoning. Understanding this shift is essential for assessing the epistemic risks that arise when scientific measurement and inference are increasingly delegated to learned, and partially opaque, instrumentation.

In this paper, I examine AI-based methods as learned measurement instruments and analyze the epistemic consequences of embedding inferential work within measurement itself. Situating AI within the broader theoretical context of scientific instrumentation clarifies what changes when measurement instruments learn and how scientific understanding can be maintained under these conditions.

Recent philosophical work has begun to examine the epistemic role of AI and machine learning in scientific practice, particularly with respect to questions of opacity, understanding,

and the distribution of epistemic labor. Some authors have explored whether AI systems challenge the traditional distinction between instruments and epistemic agents, arguing that contemporary AI may occupy intermediate or hybrid roles in scientific reasoning (Duede, 2022). Earlier work has emphasized the epistemic opacity introduced by computational methods (Humphreys, 2004) as well as the loss of contextual knowledge that accompanies the reuse of data-driven tools across domains (Leonelli, 2016). The present analysis is complementary to these approaches but differs in its point of departure: rather than asking what kind of epistemic entity AI systems are, it focuses on how the relocation of inferential work into measurement instruments reshapes measurement, uncertainty, and epistemic responsibility. On this view, the blurring of the instrument-agent boundary arises not from attributing agency to AI systems, but from structural changes in where and how inferential assumptions are enacted within scientific workflows.

In what follows, I use "AI-based methods" to refer to inferential approaches, and "AI-based systems" to refer to their concrete implementation as measurement instruments in scientific practice.

2. Classical assumptions of measurement and scientific inference

Philosophical accounts of scientific measurement have traditionally been grounded in a set of assumptions about how measurements are produced, interpreted, and justified. These assumptions are not always made explicit in scientific practice, but they structure how measurement results are understood as evidence and how epistemic responsibility is allocated (Chang, 2008; Tal, 2013). Making these assumptions explicit is necessary in order to clarify what changes when measurement instruments learn.

A central assumption of classical measurement theory is that measurement instruments implement fixed, physically grounded mappings from observables to quantities of interest. Measurement functions are typically specified in advance, justified by physical models of the measurement process, and stabilized through calibration. On this view, instruments are designed to realize well-defined relations between the world and numerical representations, and deviations from ideal measurement behavior are treated as sources of error rather than as features of the measurement function itself (Tal, 2013).

Closely related to this is a conceptual separation between measurement and inference. Measurement is understood as the production of values under controlled conditions, while inference is treated as a downstream activity through which those values are interpreted, related to existing theory, and used to support scientific claims. This division allows measurement outcomes to be treated as inputs to scientific reasoning, rather than as stages where inferential assumptions are actively negotiated. Classical accounts of measurement presuppose that while interpretation may involve judgment, the measurement process itself is not a substantive compoenent of inference in a substantive sense (Chang, 2008).

Classical measurement theory also assumes that instruments are calibratable and that measurement uncertainty can be characterised in stable and well-understood ways. Sources of uncertainty are typically modeled as stochastic noise, systematic bias, or limitations of resolution, all of which can be quantified, propagated, and reduced through improved experimental design or repeated measurement. These assumptions support the idea that

uncertainty is a property of measurement outcomes that can be managed independently of the inferential context in which measurements are later used (Hand, 2010; Tal, 2013).

Finally, traditional accounts of measurement treat interpretation and epistemic judgment as human-centered. Instruments deliver measurements, but scientists are responsible for deciding which quantities are relevant, how results should be interpreted, and whether the available evidence warrants belief or acceptance. Even when statistical or computational methods are employed, they are understood as tools that support, rather than replace, scientific judgment. On this view, epistemic responsibility attaches to scientific agents rather than to instruments or procedures.

These assumptions have shaped scientific practice across a wide range of domains. They support established practices of calibration, replication, and validation, and they sustain clear divisions of epistemic labor between instruments and scientists. However, as the following sections argue, they are increasingly strained by the introduction of AI-based measurement instruments whose behavior is shaped by learning rather than by pre-specified, physically motivated measurement functions. Understanding how and why these assumptions come under pressure is essential for assessing the epistemic consequences of AI-assisted science.

3. When measurement instruments learn

The increasing incorporation of AI into scientific measurement marks a conceptual shift in how measurement functions are constructed and justified. Classical measurement instruments implement fixed, physically motivated mappings from observables to quantities of interest. These mappings are typically specified in advance, calibrated against known standards, and evaluated in terms of well-characterized sources of error. By contrast, AI-based measurement instruments learn mappings from data, often without an explicit physical model of the underlying measurement process. This difference has important epistemic consequences.

In learned measurement systems, the relation between raw observations and measured quantities is determined through training rather than through prior specification alone. Training data, loss functions, and modeling choices become integral components of the measurement procedure, shaping what the instrument measures and how it does so. For example, in image-based measurement tasks, learned systems may infer physical properties from pixel-level data by extracting representations that are not explicitly defined in terms of underlying physical quantities. As a result, measurement outcomes depend not only on experimental conditions and calibration, but also on properties of the data used to train the system and on assumptions that may not be explicitly articulated. As a consequence, standardization is no longer secured solely through fixed measurement functions and calibration procedures, but depends on training data, modeling choices, and validation practices that may vary across contexts. The same quantity can be measured in systematically different ways across trained systems. Moreover, measurement, in this sense, is no longer separable from the inferential processes through which the mapping from data to quantities is learned.

This shift complicates traditional accounts of measurement uncertainty. Classical measurement theory typically characterizes uncertainty in terms of stochastic noise, systematic bias, or limitations of resolution, all of which can be modeled relative to fixed measurement functions. By contrast, learned measurement instruments introduce a

configuration of uncertainty that is not adequately captured by these categories alone. In the terminology standardly used in uncertainty analysis (Kiureghian and Ditlevsen, 2009), the additional uncertainty introduced by learned measurement instruments is epistemic rather than aleatory: it arises from incomplete knowledge of the measurement process itself--including model structure, training data, and inferential assumptions--rather than from irreducible variability in the measured system. In scientific contexts that rely on learned measurement instruments, such epistemic uncertainty is closely connected to limits on epistemic transparency introduced by the way inferential work is built into the measurement procedure itself. As a result, the basis on which measurement outputs are produced is no longer fully accessible to scientific users (Humphreys, 2004). While recent machine-learning literature often adopts the epistemic--aleatory distinction in operational terms, for example by associating epistemic uncertainty with model uncertainty (Kendall and Gal, 2017), the uncertainties introduced by learned measurement instruments extend beyond parameter uncertainty to include dependencies on training regimes, representational choices, and contexts of application. These uncertainties affect not only the precision of measurement outcomes, but their epistemic warrant as measurements of the relevant quantity across domains.

Attempts to address these challenges often invoke explainable AI techniques, which aim to render learned models more interpretable or transparent (Barredo Arrieta et al., 2020). While such approaches can provide valuable insights into model behavior, their applicability to scientific measurement is limited. They do not necessarily recover the inferential assumptions through which learned measurement functions are constituted. As a result, explainability may improve local intelligibility without securing the epistemic grounding of the measurement itself. Explainability methods are typically designed to highlight features or patterns associated with particular outputs, rather than to reconstruct the inferential assumptions that underpin learned measurement functions. In practice, an explanation of why a model produced a particular output may do little to clarify whether the measurement itself is valid or reliable for a given scientific purpose or domain of application. As a result, explainability alone may fall short of providing the kind of understanding required to assess the epistemic status of measurements produced by AI-based instruments.

Taken together, these developments indicate a deeper conceptual change: measurement and inference are no longer separable. When measurement instruments learn, inferential assumptions are embedded within the measurement process itself, rather than applied downstream by scientific agents. This entanglement does not imply that measurement becomes arbitrary or that scientific standards no longer apply. Rather, it signals that the epistemic role of measurement instruments has changed, and that understanding measurement now requires attention to the inferential structures through which measurement functions are learned. The point is not that learned instruments introduce new kinds of uncertainty, but that familiar forms of epistemic uncertainty are relocated into the measurement process itself, thereby altering its epistemic role.

A further consequence of this shift concerns the normative role of uncertainty in scientific practice. In classical measurement settings, uncertainty constrains inference after measurement: it qualifies the strength, scope, or reliability of claims derived from measurement outcomes, without affecting what counts as a measurement in the first place. By contrast, when measurement instruments learn, epistemic uncertainty becomes internal to the measurement process itself. Uncertainty arising from training data, model structure, and representational choices conditions how measurement outputs are stabilized, interpreted, and

treated as quantities of interest. In this sense, uncertainty no longer merely limits inference downstream, but figures in determining what counts as a measurement outcome at all. This relocation of uncertainty marks a deeper epistemic shift than the introduction of new error sources, as it alters the normative role that uncertainty plays in scientific measurement.

This transformation sets the stage for the questions addressed in the following sections. If measurement instruments participate in inferential processes that shape relevance, structure, and uncertainty, then it becomes necessary to reconsider how inferential work is distributed within scientific practice and how epistemic responsibility is assigned when measurement instruments learn. One immediate consequence of this reconfiguration is that inferential work traditionally performed by scientists is redistributed across different stages of scientific workflows, raising questions about who performs which inferential tasks, and where epistemic responsibility resides.

4. Who does the inference? Redistributing inferential work in scientific workflows

As argued in the preceding sections, the incorporation of AI into scientific measurement reshapes not only how measurements are produced, but where inferential work is carried out within scientific workflows. Inferential work is not an abstract philosophical activity, but a routine component of everyday research practice, through which measurements are transformed into evidence and scientific claims are evaluated. When parts of this work are embedded in learned measurement instruments, the distribution of inferential tasks across scientific practice is altered. Understanding this redistribution is essential for assessing the epistemic responsibilities that arise in AI-assisted science.

4.1 Inferential tasks in practice

In contemporary scientific workflows, AI-based systems are increasingly involved in tasks that go beyond the automated execution of predefined measurement procedures (Wang *et al.*, 2023). Learned models are routinely used to construct features from raw measurements, to reduce dimensionality, and to generate representations that determine which aspects of data are treated as salient. They are also used to prioritize variables or patterns for further analysis, shaping which regularities are taken to warrant explanation and which are disregarded as noise (Breiman, 2001).

In addition, AI systems are often employed to propose candidate relations or structures--such as clusters, associations, or predictive dependencies--that guide subsequent scientific inquiry. Outputs such as confidence scores, uncertainty estimates, or validation metrics further influence how results are interpreted and how much weight they are given in scientific reasoning. Although final interpretation and endorsement of claims typically remain the responsibility of human scientists, these operations affect what scientists take to be relevant evidence and how evidential support is structured prior to explicit judgment.

4.2 Shifts in epistemic roles: toolmakers and users

The redistribution of inferential work in AI-assisted science is closely tied to a growing asymmetry between the makers and users of scientific tools and systems. Scientific users often rely on computational tools and data products whose epistemic assumptions were fixed elsewhere and for other purposes, leading to a loss of contextual knowledge at the point of use (Leonelli, 2016). AI-based measurement instruments are often designed, trained, and validated by method specialists who are not domain experts in the scientific phenomena to which the tools are later applied. For example, machine-learning models for medical image analysis are often developed and validated in computer science-led research contexts, with clinical expertise contributing primarily through data annotation and evaluation, and are later adopted by clinicians or biomedical researchers as measurement instruments without direct access to the full range of training data, modeling assumptions, or systematic knowledge of failure modes outside the original validation context (Litjens *et al.*, 2017). Scientific users, in turn, adopt these systems as components of measurement workflows, frequently without direct access to training data, modeling assumptions, or detailed knowledge of failure modes. In this way, scientific users rely on tools developed by others, creating forms of epistemic dependence (Hardwig, 1985) that are sustained by shared epistemic norms governing trust, authority, and responsibility within scientific practice (Pihlström, 2020). The link to normativity here concerns the structuring of epistemic roles and responsibilities in scientific practice, not the justification of specific measurement outcomes.

In such contexts, structured by shared epistemic norms governing trust, authority, and responsibility, reliance on AI-based instruments--where they are adopted--becomes a practical necessity in scientific practice rather than an epistemically transparent choice. Trust in measurement outputs is often grounded in institutional reputation (Leonelli, 2016), prior use, or reported performance metrics, rather than in an understanding of how inferential assumptions are enacted within the system. This shift reframes agency not in terms of autonomy or intentionality, but as a displacement of epistemic roles within scientific workflows, in which key inferential functions are carried out upstream of scientific interpretation.

4.3 Why this blurs the boundary between instrument and agent

Traditionally, scientific instruments have been understood as implementing measurement procedures defined and interpreted within established conceptual frameworks, while inference and judgment were regarded as the responsibility of scientific agents. Learned measurement instruments complicate this division. By performing operations that shape relevance, structure evidential relations, and characterize uncertainty, they participate in inferential processes that historically belonged to human scientific reasoning.

This blurring of the boundary between instruments and agents (Baird, 2004) should not be understood as a claim about AI systems possessing intentions, understanding, or agency in a strong sense. Rather, it reflects a shift in where epistemic decisions are enacted within scientific practice. When inferential work is embedded in measurement instruments, the distinction between producing measurements and shaping evidence becomes less clear, even though responsibility for interpretation and endorsement remains human.

4.4 Implications for scientific accountability

As inferential work is increasingly embedded in measurement instruments, there is a risk that evaluative output metrics--such as confidence scores, validation metrics, or performance benchmarks--come to stand in for epistemic judgment, rather than informing it. When such metrics are treated as sufficient grounds for acceptance, the normative act of judging whether results warrant belief or use--and of taking responsibility for that judgment--may be obscured rather than supported.

These challenges are further complicated by forms of epistemic opacity that arise when inferential work is embedded in computational systems. Opacity is not merely a matter of insufficient transparency, but a structural feature of computationally extended science (Humphreys, 2004). In the case of learned measurement instruments, this opacity complicates efforts to locate responsibility for inferential assumptions and to assess the epistemic status of measurement outputs.

This shift has direct implications for scientific accountability. When inferential work is embedded in instruments, responsibility for errors or biases may originate upstream of scientific interpretation, in choices related to data selection, model design, or training regimes. Traditional epistemic safeguards, such as replication or calibration, remain essential but may be insufficient without an understanding of the learned components of measurement instruments.

The significance of this shift lies not in the increasing power or apparent autonomy of AI systems, but in the structural delegation of inferential work into measurement instruments--a change with important epistemic consequences. Recognizing this delegation sets the stage for analyzing the epistemic risks and responsibilities that arise in AI-assisted science.

Section 5. Epistemic risks and responsibilities in AI-assisted science

The epistemic risks associated with AI-assisted science do not arise from the power, autonomy, or perceived mystery of AI systems. Rather, they follow from a structural transformation in scientific practice: the embedding of inferential work within measurement instruments themselves. When measurement instruments learn, they do not merely automate existing procedures; they delegate inferential assumptions that were traditionally enacted through scientific reasoning into the operation of instruments. This delegation has consequences for how evidence is generated, interpreted, and trusted.

This concluding section articulates how this shift affects epistemic risk and responsibility in scientific practice. The risks at issue are not primarily ethical or social, but epistemic: they concern the conditions under which measurements can function as evidence, the transparency of inferential assumptions, and the attribution of responsibility for scientific claims. Understanding these risks is also necessary for clarifying how scientific understanding can be sustained when using learned measurement instruments that are often only partially non-transparent.

5.1 Epistemic risks of learned measurement

Learned measurement instruments introduce a distinctive class of epistemic risks that arise from their structure rather than from their performance alone. These risks are not reducible to increased complexity or to the use of statistical models, but follow from the fact that inferential work is embedded within the measurement process itself.

One such risk is epistemic opacity. When measurement functions are learned from data, the inferential steps that connect raw observations to measured quantities are often inaccessible to scientific users. This opacity is not merely a matter of insufficient documentation or explanation, but reflects the way inferential assumptions are encoded across training data, model architectures, and optimization procedures. As a result, scientists may lack the means to assess which features of the data are driving measurement outcomes or how these relate to underlying physical or theoretical quantities.

A second risk concerns dependence on training data and modeling choices. Because learned measurement instruments acquire their behavior through training, their epistemic reliability is inseparable from properties of the datasets and modeling decisions that shaped them. Choices about data selection, labeling, preprocessing, and loss functions effectively become part of the measurement procedure, even though they may be invisible at the point of use. This dependence creates vulnerabilities when instruments are applied outside the conditions under which their inferential assumptions were established.

Relatedly, learned measurement instruments can be fragile under domain shift. Measurement systems that perform reliably under one set of experimental or observational conditions may yield systematically distorted results when applied to data drawn from different regimes. Such fragility is epistemically significant because it affects not only the accuracy of measurements, but their interpretability and scope. When domain shifts are subtle or poorly characterized, measurement outputs may retain an appearance of reliability even as their epistemic grounding erodes.

Finally, learned measurement instruments often compress complex forms of uncertainty into simplified output metrics. Confidence scores, validation measures, or performance benchmarks can provide useful summaries of model behavior, but they may obscure the sources and structure of uncertainty that are epistemically relevant to scientific judgment. When uncertainty is reduced to a single number, distinctions between noise, model inadequacy, and contextual mismatch may be lost, increasing the risk that such metrics are treated as substitutes for judgment rather than as inputs to it.

Taken together, these risks arise not from the use of AI as such, but from the structural features of learned measurement. Recognizing their origin is a necessary step toward understanding how epistemic responsibility is redistributed in AI-assisted science, and how scientific understanding might be preserved when measurement instruments learn.

5.2 Limits of classical epistemic safeguards

Classical epistemic safeguards such as calibration, validation, replication, and peer review have long played a central role in securing the reliability of scientific measurements. These practices remain essential in AI-assisted science, but their effectiveness presupposes

conditions that are increasingly strained when inferential assumptions are embedded in learned measurement instruments. Understanding these limits is crucial for assessing how epistemic responsibility can be maintained under conditions of learned measurement.

Calibration and validation traditionally presuppose stable measurement functions whose behavior can be characterized independently of particular datasets. Calibration procedures assume that deviations from ideal measurement can be identified and corrected through comparison with known standards, while validation assesses performance under specified conditions. Learned measurement instruments complicate these assumptions. When measurement functions are shaped by training data and optimization objectives, calibration and validation may capture performance only relative to specific data distributions, leaving their behavior under other conditions underdetermined.

Replication faces related challenges. Replication presupposes that comparable instruments, operating under sufficiently similar conditions, will produce comparable results. In the case of learned measurement instruments, nominally identical systems may differ in training data, preprocessing steps, or model updates, even when their external specifications are the same. As a result, failures of replication may reflect differences in embedded inferential assumptions rather than experimental error or theoretical inadequacy. This complicates the interpretation of both successful and unsuccessful replication attempts.

Classical error models also assume that uncertainty can be represented in interpretable and stable forms, typically as noise or bias associated with measurement outcomes. Learned measurement instruments often compress multiple sources of uncertainty--arising from data limitations, model structure, and contextual mismatch--into simplified metrics that do not transparently map onto these classical categories. When uncertainty is not interpretable in familiar terms, error propagation and uncertainty management become less reliable guides to epistemic judgment.

Peer review remains a crucial safeguard, but it too is affected by these shifts. Reviewers can assess experimental design, theoretical coherence, and reported performance metrics, yet they may lack access to the inferential assumptions embedded in learned measurement instruments or the expertise required to evaluate them. As a result, peer review may inadvertently reinforce trust in outputs without fully interrogating the conditions under which those outputs are epistemically warranted.

Similar concerns have been raised in earlier discussions of rule-based scientific practice, where attempts to replace judgment with standardized procedures were shown to redistribute rather than eliminate epistemic responsibility (Daston 2022).

Taken together, these considerations do not undermine the value of classical epistemic safeguards. Rather, they indicate that such safeguards are no longer sufficient on their own to secure scientific understanding when measurement instruments learn. Recognizing their limits is a necessary step toward clarifying how epistemic responsibility must be redistributed and how scientific practice can adapt to the structural changes introduced by AI-assisted measurement.

5.3 Distributed epistemic responsibility

When inferential work is embedded in measurement instruments, epistemic responsibility becomes distributed across scientific practice. Decisions that shape what counts as relevant evidence, how uncertainty is characterized, and which relations are emphasized are no longer located solely in human interpretation, but are partly enacted within learned measurement systems. As a result, responsibility for scientific claims cannot be straightforwardly attributed to a single actor or stage of inquiry, but is spread across tool design, data curation, modeling choices, and scientific use.

A particular epistemic risk arises when outputs of learned instruments--such as confidence scores, validation metrics, or performance benchmarks--are treated as substitutes for epistemic judgment rather than as inputs to it. While such metrics summarize aspects of model behavior or empirical performance, they do not by themselves determine whether a result warrants belief, acceptance, or use in scientific reasoning. Epistemic judgment involves a normative assessment of evidential support, scope, and limitation, and it remains a responsibility of scientific agents rather than of instruments.

The redistribution of epistemic responsibility in AI-assisted science has a partial analogue in the rise of large, multi-author scientific collaborations, in which epistemic labor is distributed across many contributors. In such cases, however, responsibility remains anchored in human agents who possess epistemic judgment and can, at least in principle, justify and defend the claims to which they attach their names. By contrast, learned measurement instruments may participate in inferential processes without possessing such judgment, making the distribution of responsibility less explicit and more difficult to trace.

The risk of substituting output metrics for epistemic judgment is not entirely new in scientific practice. Similar concerns have been raised in earlier debates about statistical significance testing, where procedural thresholds were criticized for standing in for epistemic judgment rather than supporting it (Gigerenzer, 2004). What distinguishes AI-assisted measurement is that inferential assumptions are increasingly embedded within instruments themselves, shaping measurement outcomes prior to explicit judgment.

Delegating epistemic judgment to learned measurement systems does not transfer responsibility; it obscures it. When judgment is replaced by output metrics, responsibility for endorsing scientific claims becomes difficult to locate, even though it has not disappeared. This risk is amplified in contexts where AI-based measurement instruments are developed by toolmakers who are not domain experts, and where scientific users rely on these instruments without full access to their assumptions, training data, or conditions of validity.

Recognizing this distribution of epistemic responsibility is essential for understanding the epistemic consequences of AI-assisted science. Learned measurement instruments can inform scientific judgment by providing structured summaries of data and uncertainty, but they cannot assume responsibility for endorsing claims. Maintaining scientific understanding therefore requires preserving a clear distinction between inferential outputs and the epistemic judgments through which those outputs are evaluated and integrated into scientific knowledge.

5.4 Maintaining scientific understanding

If scientific understanding is to be maintained in the presence of learned measurement instruments, it must be reconceived in light of the delegation of inferential work into instruments themselves. The question is not whether AI systems should be used in scientific practice, but under what conceptual conditions their use can support understanding rather than merely produce reliable outputs. Addressing this question requires clarity about the epistemic role of learned instruments and the responsibilities that remain with scientific agents.

A first requirement is recognizing learned measurement instruments as participants in inferential processes, rather than as neutral channels through which measurements pass unchanged. This recognition does not attribute agency or understanding to instruments, but it acknowledges that they enact inferential assumptions that shape relevance, structure evidence, and condition uncertainty. Treating learned instruments as purely passive tools obscures their epistemic contribution and makes it more difficult to assess how measurements function as evidence.

A second requirement is that the scope, assumptions, and limitations of learned measurement instruments be made explicit to the extent necessary for epistemic assessment. Scientific understanding depends not only on knowing what an instrument outputs, but on knowing the conditions under which those outputs can be treated as meaningful measurements. When inferential assumptions are embedded in training data and model design, understanding requires access--conceptual if not technical--to how those assumptions constrain the validity and applicability of measurement results.

Finally, maintaining scientific understanding requires preserving human responsibility for explanation and interpretation. Recent discussions of AI-assisted science likewise emphasize that computational systems can contribute to scientific understanding without themselves being epistemic agents, provided that explanation and interpretation remain grounded in human reasoning (Krenn *et al.*, 2022). While learned instruments may generate measurements, summaries, or candidate relations, they do not themselves explain phenomena or justify scientific claims. Explanation involves situating results within theoretical frameworks, articulating causal or structural relations, and assessing the adequacy of inferential support--activities that remain the responsibility of scientific agents, even as the tools they use become more inferentially complex.

Taken together, these requirements clarify what it means to sustain scientific understanding under conditions of learned measurement. Understanding is preserved not by resisting AI-assisted methods, but by maintaining clear epistemic roles and responsibilities as inferential work is redistributed. Recognizing learned instruments as inferential participants, while retaining human judgment and responsibility, is essential for ensuring that scientific knowledge remains intelligible and accountable when measurement instruments learn.

5.5 Concluding position

The growing use of artificial intelligence in science marks a shift not only in how measurements are performed, but in where inferential work is carried out. When measurement instruments learn, parts of the reasoning that connect observations to scientific

claims are no longer located exclusively in human judgment, but are embedded within computational systems that shape relevance, structure, and uncertainty before interpretation begins. This delegation of inferential work alters the epistemic role of measurement without eliminating the need for scientific reasoning.

The epistemic risks associated with AI-assisted science arise from this transformation of measurement itself. Learned instruments introduce dependencies on training data, modeling choices, and implicit assumptions that are not always transparent to scientific users and that are not fully addressed by classical accounts of calibration, validation, or error. These risks do not stem from AI systems acting autonomously or possessing agency in a strong sense, but from the way learned measurement entangles inference with instrumentation.

Recognizing this shift is essential for sustaining scientific understanding in the presence of AI. Scientific knowledge can be maintained not by treating AI systems as neutral tools or independent agents, but by acknowledging learned measurement instruments as participants in inferential processes whose assumptions and limitations must remain subject to critical scrutiny. When measurement instruments learn, preserving scientific understanding requires resisting the substitution of output metrics for epistemic judgment, and recognizing that responsibility for endorsing scientific claims cannot be delegated without loss.

Acknowledgements

AI-based language tools were used to assist with drafting and revising portions of this manuscript. The author retains full responsibility for the content, interpretation, and conclusions. No empirical measurements were conducted.

References

Baird, D. (2004) *Thing Knowledge*. University of California Press.

Barredo Arrieta, A. *et al.* (2020) “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, 58, pp. 82–115. Available at: https://doi.org/10.1016/j.inffus.2019.12.012.

Breiman, L. (2001) “Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author),” *Statistical Science*, 16(3), pp. 199–231. Available at: https://doi.org/10.1214/ss/1009213726.

Chang, H. (2008) *Inventing Temperature: Measurement and Scientific Progress*. Oxford, New York: Oxford University Press (Oxford Studies in Philosophy of Science).

Duede, E. (2022) “Instruments, agents, and artificial intelligence: novel epistemic categories of reliability,” *Synthese*, 200(6), p. 491. Available at: https://doi.org/10.1007/s11229-022-03975-6.

Gigerenzer, G. (2004) “Mindless statistics,” *The Journal of Socio-Economics*, 33(5), pp. 587–606. Available at: https://doi.org/10.1016/j.socec.2004.09.033.

Hand, D. (2010) *Measurement Theory and Practice: The World Through Quantification*. Wiley.

Hardwig, J. (1985) "Epistemic Dependence," *The Journal of Philosophy*, 82(7), pp. 335–349. Available at: https://doi.org/10.2307/2026523.

Humphreys, P. (2004) *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford, New York: Oxford University Press.

Kendall, A. and Gal, Y. (2017) "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (NIPS'17), pp. 5580–5590. Available at: https://dl.acm.org/doi/10.5555/3295222.3295309 (Accessed: January 16, 2026).

Kiureghian, A.D. and Ditlevsen, O. (2009) "Aleatory or epistemic? Does it matter?," *Structural Safety*, 31(2), pp. 105–112. Available at: https://doi.org/10.1016/j.strusafe.2008.06.020.

Krenn, M. *et al.* (2022) "On scientific understanding with artificial intelligence," *Nature Reviews Physics*, 4(12), pp. 761–769. Available at: https://doi.org/10.1038/s42254-022-00518-3.

Leonelli, S. (2016) *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.

Litjens, G. *et al.* (2017) "A survey on deep learning in medical image analysis," *Medical Image Analysis*, 42, pp. 60–88. Available at: https://doi.org/10.1016/j.media.2017.07.005.

Pihlström, S. (2020) "How Is Normativity Possible? A Holistic-Pragmatist Perspective," in I. Niiniluoto and S. Pihlström (eds.) *Normativity*. Helsinki: Philosophical Society of Finland (Acta Philosophica Fennica), pp. 201–228.

Tal, E. (2013) "Old and New Problems in Philosophy of Measurement," *Philosophy Compass*, 8(12), pp. 1159–1173. Available at: https://doi.org/10.1111/phc3.12089.

Wang, H. *et al.* (2023) "Scientific discovery in the age of artificial intelligence," *Nature*, 620(7972), pp. 47–60. Available at: https://doi.org/10.1038/s41586-023-06221-2.